Statistics in Medicine

Reference bias in reports of drug trials

PETER C GØTZSCHE

Abstract

Articles published before 1985 describing double blind trials of two or more non-steroidal anti-inflammatory drugs in rheumatoid arthritis were examined to see whether there was any bias in the references they cited. Altogether 244 articles meeting the criteria were found through a Medline search and through examining the reference lists of the articles retrieved. The drugs compared in the studies were classified as new or as control drugs and the outcome of the trial as positive or not positive. The reference lists of all papers with references to other trials on the new drug were then examined for reference bias. Positive bias was judged to have occurred if the reference list contained a higher proportion of references with a positive outcome for that drug than among all the articles assumed to have been available to the authors (those published more than two years earlier than the index article). Altogether 133 of the 244 articles were excluded for various reasons—for example, 44 because of multiple publication and 19 because they had no references. Among the 111 articles analysed bias was not possible in the references of 35 (because all the references gave the same outcome); 10 had a neutral selection of references, 22 a negative selection, and 44 a positive selection—a significant positive bias. This bias was not caused by better scientific standing of the cited articles over the uncited ones.

Thus retrieving literature by scanning reference lists may produce a biased sample of articles, and reference bias may also render the conclusions of an article less reliable.

Introduction

In literature retrieval data searches may be insufficient[1 2] and often have to be supplemented by scanning the reference lists. This may lead to a biased selection of articles, particularly if the reference lists reflect the authors' prejudices. The existence of a one sided reference bias was suggested by Sackett in 1979 concerning two articles on the inheritance of hypertension,[3] but it seems not to have been shown statistically.

Reports of drug trials are an excellent opportunity for studying this possible bias: they may be numerous and the outcome can often be easily classified. This study examined trials of non-steroidal anti-inflammatory drugs.

Methods

I collected articles published before 1985 on double blind trials of two or more of the 17 non-steroidal anti-inflammatory drugs marketed in Denmark, given in repeated doses as tablets or capsules to patients with rheumatoid arthritis. For multiple publications on the same patients the article with the largest number of references was analysed. I excluded trials on several diseases if the outcome could not be evaluated for rheumatoid arthritis alone and trials published solely as abstracts.

I performed a Medline search covering 1966 and onwards in May 1985. One or more of the drugs and *arthritis, rheumatoid* were combined as main headings with the boolean operator "and" and further combined with either *comparative study, review,* or *dose-response relationship, drug* (introduced in 1973), also as main headings.[4] When they were not indexed as main headings I searched for the drugs as text words. I also contacted the companies marketing the proprietary preparations, and, finally, I scanned the reference lists of the collected articles.

I read the articles in random order chosen according to a table of random numbers. For each article the drug that seemed to be the authors' primary interest was labelled the "new" drug. This was usually evident from the title, the introduction, supply of coded drugs, grants, and statistical advice. Other drugs were defined as "control" drugs. Thereafter, I read the authors' conclusion and took it at face value. The outcome for any drug was defined as "positive" if (*a*) it had the same effect as another with less pronounced side effects, (*b*) it had a better effect without more pronounced side effects, or (*c*) it was preferred more often by the patients when the effect and side effect evaluation were combined. If none of these criteria was fulfilled, or if any differences were considered unimportant by the authors (whether the differences were statistically significant or not was immaterial), the outcome for the drug was "not positive."

Thus the drugs in each article were classified as new or control, and the outcome for each drug was independently classified as positive or not positive.

After classification I examined the reference lists of all papers with references to other double blind trials in rheumatoid arthritis on the new drug for reference bias. For each article I noted whether the proportion of references to trials with a positive outcome for the new drug was higher, the same, or lower than the proportion among all articles assumed to have been available to the authors at the time of submitting the manuscript—that is, apart from those referred to by themselves, all other articles on the new drug published two years or more before the article examined.

For example, an article comparing ketoprofen as new with aspirin as the control might refer to three trials with ketoprofen, of which two had a positive outcome and one a not positive outcome (whether ketoprofen was positive compared with aspirin in the article examined and whether ketoprofen was new or control in the references were unimportant). If there were two additional trials of ketoprofen, both published at least two years previously and both with a not positive outcome for ketoprofen, then a positive selection of references on ketoprofen would be shown for the article examined, because the proportion of trials with a positive outcome in the reference list, two out of three, was higher than that among the available trials, two out of five. In general, the selection of references might be positive, neutral, or negative, according to the sign of the difference between the two proportions.

I repeated the analysis, considering only references in English.

Since one reason for not referring to some trials might have been that the control drugs were not marketed in the authors' own country I repeated the analyses in a modified form. For this I restricted the available references to trials on the controls chosen by the authors themselves—that is, as they appeared in their article and in its references (authors' control drugs). Thus an article with ketoprofen as new and aspirin as control which also referred to two articles comparing ketoprofen with indomethacin would correspond to a restricted reference sample of these two articles together with all the others comparing ketoprofen with either aspirin or indomethacin published at least two years previously. By contrast, the sample for the main analysis would consist of the two articles plus all others on ketoprofen which were at least two years old.

The possible influence of the scientific standing of the journals was studied

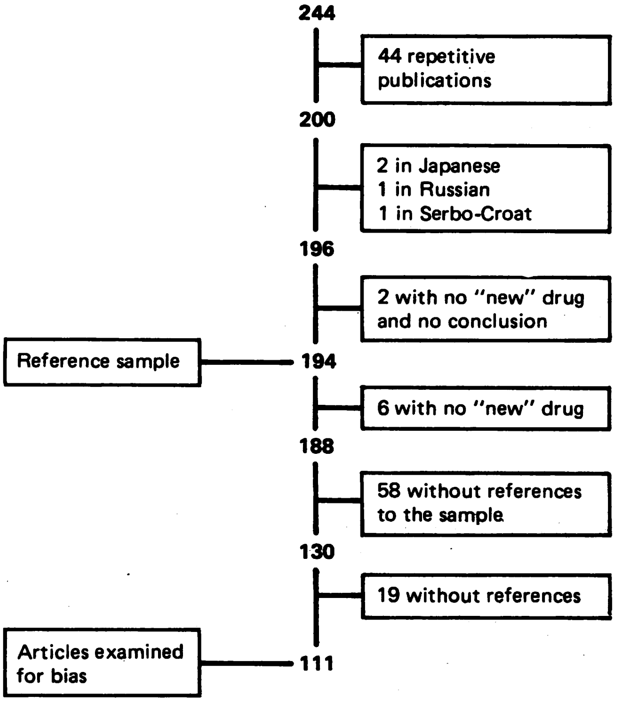
Liver Unit, Hvidovre Hospital, DK-2650 Hvidovre, Denmark
PETER C GØTZSCHE, MD, MSC, registrar

by using the *Science Citation Index*.[5] I subdivided the choice of references in the main analysis by whether the rank sum of the cited journals was lower, the same, or higher than the proportional rank sum of those not cited (the most cited journals, with the highest impact factors, were given the lowest ranks, and journals not indexed the highest).

Under the null hypothesis of no reference bias, articles with a positive selection of references on the new drug should have been expected to appear with the same frequency as articles with a negative selection. Ignoring articles with a neutral selection, I used a sign test.

Results

Exclusion of trials—A total of 244 articles was collected. After thorough cross checking by drugs, coauthors, and results 43 of these were found to be multiple publications, and one was strongly suspected of being so and excluded (figure). The fact that these had been published elsewhere was noted in only 12 of these articles. The results of 20 trials were published twice, 10 trials three times, and one trial five times, usually in full. Of the remaining 200 articles, 155 were in English. Two articles in Japanese, one in Russian, and one in Serbo-Croat had to be excluded from my analysis, since neither the companies nor the university library had translations; no article used them as references. Two multiple drug trials were excluded, since no new drug could be identified and they had no conclusion. The remaining 194 articles were included in the reference sample (figure).


244
44 repetitive publications
200
2 in Japanese
1 in Russian
1 in Serbo-Croat
196
2 with no "new" drug and no conclusion
Reference sample
194
6 with no "new" drug
188
58 without references to the sample
130
19 without references
Articles examined for bias
111

Definition of the sample studied; figures are number of articles.

Articles lacking references—In six of the 194 articles a new drug could not be defined and these could not be assessed for bias. Nor was this possible in a further 77 articles, 58 of which had no references to other double blind trials and 19 no references at all (figure). This low level of citation was also found when only the drugs which the authors had studied were considered. Reference to other trials with the same new and control drug was omitted in 38% of articles in which it would have been possible. The omissions were of the same order of magnitude when a three year limit for the availability of the references was used.

Reference bias—Table I shows the analysis of the remaining 111 articles. The number of articles in which the available references were exclusively positive or not positive is shown separately, since no bias was possible in these cases. There was a significant bias towards a positive selection of references on the new drug, both for any language (p<0·01) and for references in English (p<0·05). For any language, when the 35 articles in which bias was impossible and the 10 with a neutral selection were ignored then 44 of the remaining 66 articles (67%; 95% confidence interval 54% to 78%) had an overrepresentation of references to trials with a positive outcome for the new drug. With a three year limit for the references the bias was also apparent (p<0·01 for both analyses). The bias was not caused by overrepresentation of highly cited journals among the articles with a positive selection of references (table II).

TABLE I—*Number of articles with positive, neutral, and negative selection of references to the drug of primary interest. One article had no references in English*

	Positive selection	Neutral selection	Negative selection	Bias not possible	Total	p
All control drugs:						
Any language	44	10	22	35	111	<0·01
English only	38	8	20	44	110	<0·05
Authors' control drugs:						
Any language	30	9	19	53	111	<0·20
English only	26	9	18	57	110	<0·30

TABLE II—*Number of articles with positive, neutral, and negative selection of references to the drug of primary interest, divided as to whether the rank sum of the cited journals according to the Science Citation Index*[5] *was lower than, equal to, or higher than the proportional rank sum of those not cited. The most cited journals (with highest impact factors) were given the lowest ranks*

	Positive selection			Neutral selection			Negative selection		
	Low	Equal	High	Low	Equal	High	Low	Equal	High
All control drugs:									
Any language	22	3	19	2	7	1	12	2	8
English only	15	6	17	0	8	0	12	3	5

Exclusion of 13 articles that could not be provided through university libraries in Scandinavia, Britain, or Germany but were obtained from the companies—thereby giving them the same status as unpublished reports—did not eliminate the bias (p<0·05 for both analyses).

The analysis for the authors' control drugs showed a similar, but non-significant, bias (table I). There was no trend towards a positive selection of articles in the Medline search or in the lists provided by the companies on their own drug.

Discussion

This study has shown a high frequency of multiple publication and reference bias. Multiple publication was sometimes difficult to detect: the number of authors might differ; the first author might vary; the title might be different; and the name of the institution might be omitted. If multiple publication is not detected it may cause problems in any pooled analysis of trials (meta-analysis) or mislead the reader of the individual article. In fact, five articles referred to multiple publications by others as if these concerned separate trials. Multiple publication was often due to company sponsored symposia, published as supplements, and the motive was not apparently to have versions in different languages, since these were different in only 12 of the 44 articles. Nevertheless, multiple publication was also frequently seen in current journals. One trial was published twice in the same journal, with 104 patients initially and six patients added five years later, without any reference to the first article.

The larger the number of possible references the less should be the impact of being unaware of a single reference. Thus any reference bias should be at its greatest when the authors have many articles to choose from. This was exactly the case: the bias for eight or more possible references was statistically significant (p<0·01, table III). This finding might explain why the analysis for the authors' control drugs showed a non-significant bias (table I);

TABLE III—*Number of articles with positive, neutral, and negative selection of references to the drug of primary interest in relation to number of possible references. Any control drug, any language (n=111)*

	Positive selection	Neutral selection	Negative selection	Bias not possible	Total
1-3 possible references	5	5	4	26	40
4-7 possible references	16	3	11	8	38
≥8 possible references	23	2	7	1	33

fewer references are possible in any restricted analysis while, correspondingly, the number of articles where no bias is possible will be higher.

A manual search of journals might have identified some further articles, but I did not know which journals to look in. The 200 articles were published in 63 journals or journal supplements, as well as in a few symposia in book form. Even so, given that I made great efforts to secure as complete a sample as possible, using standard methods, I believe that any undetected articles would have been unlikely to affect the results of this study.

A decision to refer to a particular trial may well depend on the quality of the methods used, and hence I analysed only double blind trials. (Such studies are usually also randomised, thus fulfilling what are probably the two most important methodological criteria for clinical trials.) Surprisingly, many articles had no references to other double blind trials with the same drugs. Thus, the reference pattern was somewhat irrelevant, since the aim of these trials, all with aı. active control drug, was pragmatic, trying to solve the question of which drug should be preferred.[6]

The trials that were least cited in the reference lists were not published in journals or books that are difficult to obtain either in the library or through a Medline search, nor did they concern unfamiliar drugs. In fact, the reference bias was caused mainly by a biased selection of references on indomethacin, the most common control drug used in the study. Reference was made only twice to trials on controls not represented in the sample. A bias in the initial classification of drugs as positive or not positive is unlikely, since it would have been impossible to foresee what given judgments would have led to in the analysis, carried out months later.

In conclusion, therefore, the reference bias shown in this study seems to be real. Such a finding has important implications, since there is no reason to believe that rheumatologists are more biased than others in selecting references. A reader tracing the literature on any new drug using the reference lists given in the articles might risk obtaining a biased sample. Reference bias has another serious implication: it may render the conclusion of the individual article less reliable. Is this also true for review articles, and for other disciplines in medicine?

The study was supported by a grant from the Danish Medical Research Council. I thank the University Library II, Copenhagen, the medical companies, and Alice Nørhede, librarian at Herlev Hospital, for help in data collection; Dr John Anderson for linguistic help; and, especially, Dr Thorkild I A Sørensen, liver unit, Hvidovre Hospital, for his valuable suggestions and comments on the manuscript.

References

1 Poynard T, Conn HO. The retrieval of randomized clinical trials in liver disease from the medical literature: a comparison of MEDLARS and manual methods. *Controlled Clin Trials* 1985;6:271-9.
2 Dickersin K, Hewitt P, Mutch L, Chalmers I, Chalmers TC. Perusing the literature: comparison of MEDLINE searching with a perinatal trials database. *Controlled Clin Trials* 1985;6:306-17.
3 Sackett DL. Bias in analytic research. *J Chronic Dis* 1979;32:51-63.
4 National Library of Medicine. *Medical subject headings: annotated alphabetic list.* Bethesda, Maryland: NLM, 1985.
5 Institute for Scientific Information. *Science citation index. Journal citation reports.* Philadelphia: ISI, 1986.
6 Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chronic Dis* 1967;20:637-48.

Towards a reduction in publication bias

ROBERT G NEWCOMBE

Abstract

Current practice results in the publication of many research studies in medical and related disciplines which may be criticised on the grounds of inadequate sample size and statistical power. Small studies continue to be carried out with little more than a blind hope of showing the desired effect. Nevertheless, papers based on such work are submitted for publication, especially if the results turn out to be statistically significant. There is confusion about what makes a result suitable for publication. Often there is a preference for statistically significant results at the peer review stage. Consequently published reports of small studies tend to contain too many false positive results and to exaggerate the true effects.

The use of a criterion of a posteriori power does not eliminate the bias; a priori power is the criterion of choice. This could be implemented by peer review of study protocols at the planning stage by funding bodies and journals.

Introduction

Profound biological and behavioural differences between human beings mean that statistical methods have to be used in presenting medical research findings in an unbiased way. Hence statisticians have devised methods of estimation and significance testing, which are now widely used. Nevertheless, though the mathematical aspects of these methods are acceptable, what is done with the results commonly leads to serious selection bias. An article that reports a statistically significant difference between two treatments is more likely to be published than one which does not. Many research studies have inadequate numbers of subjects, and significance can be attained only if chance conveniently exaggerates the difference.

So long as statistical significance is used as a major criterion of acceptability for publication the published results of medical research will contain a high proportion of false positive results. Thus quantitative estimates of treatment effects taken from published work cannot be regarded as free from bias. There are established methods to calculate the power of a study, which is the probability of detecting a specified, important difference using a test with a set significance level. The interpretation of statistical power is satisfactory only when it is calculated with values specified at the design stage of the study. The proper method to assess the adequacy of the sample size is by peer review of values specified in the protocol. If this is done the significance level eventually attained is no longer relevant to selection for publication.

Department of Medical Computing and Statistics, University of Wales College of Medicine, Cardiff CF4 4XN
ROBERT G NEWCOMBE, MA, PHD, lecturer in medical statistics

Importance of sample size

Manuscripts submitted to medical journals often contain serious statistical faults.[1] Various steps have been taken to remedy this,